单元 4 设计物种分布模型

欢迎大家回到物种分布模型的在线开放课程。

现在我们对物种分布模型的理论背景和建模所需要的不同类型数据有了更多了解，我们可以进行模型设计了。你需要考虑你要解决的科学问题，以及需要哪种数据和算法来解决。正如我在上一个单元中提到的，俗话说，“输入垃圾，输出也会是垃圾”，如果你只是随意地把一些数据输入模型，结果会没有什么意义。网络上公开的大量数据和虚拟实验室等工具使得我们可以轻松运行物种分布模型，最好，你在运行模型前，先评估输入数据的质量，以确保你的结果可靠。

为了设计物种分布模型，你需要仔细考虑模型的三个组成部分：物种数据，环境数据和算法。我们提到的算法是指基于一组环境数据来确定物种分布概率的实际方法。哪种算法最合适则取决于你的数据类型，当然反过来也可以：在你确定使用的算法基础上选择数据，比如有多种最佳方法可以用来生成伪分布无数据。

我们从几个一般性问题开始，回答你需要什么样数据的问题：首先，你感兴趣的物种有哪些数据是可以获取到的，这些数据是否准确？正如我们在课程的上一个单元中所提到的，检查物种分布的数据集中是否存在异常值非常重要，或者，考虑是否存在任何取样偏差，如数据覆盖的地理范围。当然，对环境数据也要做同样的检查。

你可能想知道物种分布记录的数据量需要有多大，这意味着有较好的模型运行结果时需要多少个物种分布点。当然，这取决于已有数据量。对于一些常见的物种，如楔尾雕，你可能会找到数以万计的分布记录。但出于物种保护的目的，你可能对一个记录少得多的罕见物种，如里士满青蛙感兴趣。你需要的最佳分布记录数量与物种分布范围的大小有关。一般来说，与地理分布范围较小和环境忍耐度有限的物种相比，地理分布范围广并且对一系列环境条件忍耐度较大的物种的模型预测结果的准确性往往相对较低。所以即使你的物种罕见到只有几个记录，如果其地理分布范围很小，其样本的环境条件可能相比分布广的物种，更能准确代表它的适宜生境。通常来说，物种分布位点数量不能少于 30，仅使用物种分布数据的算法受小样本量的影响较小。

在你开始搜集环境数据之前，你应该考虑哪些因素可能会影响你感兴趣的物种的分布。虽然一些算法能够处理大量的预测变量，但仍然应该严格筛选放入模型的变量。这意味着你必须做一些研究来了解物种，并选择直接影响物种分布的预测因子。例如，如果你知道你的物种对极高温或极低温敏感，那么确保模型包含温度相关变量。如果你不确定哪些因素影响你的物种，可以先运行一个有大量预测因子的模型。模型的结果会显示每个环境变量的响应曲线，可以用它作为筛选依据，来选择最重要的预测因子来运行后续更精细的模型。在这个例子中，土壤类型和辐射的响应曲线显示为一条水平直线，这意味着它们对物种的分布概率没有影响，因此在下一个模型中可以选择剔除这些变量。你必须意识到，大多数算法都考虑到变量之间的交互作用，因此添加或删除变量会改变模型结果。这再次突显了设计物种分布模型时进行一些研究的重要性。

物种分布模型的第三个方面，你必须选择的是用于关联物种分布数据与环境条件的算法。有很多不同的算法可用于物种分布模型。在这个课程中，我们重点关注四类模型：地理模型，框架模型，统计回归模型和机器学习模型。这种分类并非一成不变，而且有点随意，因为许多机器学习模型都是基于回归技术，同时也用于统计回归模型。所以你可能在其他地方看到不同的模型分类方法，在这里，我将使用这个分类来快速概述我们将在本课程的第 5、6、7 单元中详细讨论的算法。

地理模型只使用物种分布数据，而不使用环境数据。它们在一定地理空间内运行，所以可以在坐标轴上进行地图上的可视化展示。这些模型使用简单的算法，预测物种在分布位点周围特定形状或特定距离范围内的分布。所以在这个例子中，模型围绕最外面的分布点绘制了一个边界，预测该物种可存在于该边界内的任何地方，这里用绿色表示。因为地理模式不考虑分布地点的环境条件，被认为不是真实的物种分布模型。但它们提供了一个很好的方法来快速了解一个物种的空间分布范围。

框架模型是最基本的预测物种分布的模型。和地理模型一样，它们可以只使用物种分布数据，但也可以使用环境数据。因此，它们在环境空间中运行，图上的轴表示用于预测物种分布概率的不同环境变量。最著名的框架模型是 Bioclim 模型，它被认为是第一个物种分布模型。Bioclim 根据每个环境变量的最小值和最大值构建一个边界框，只要是在边界内，物种便可能有分布。框架模型有一些局限性，它们只能处理连续的环境变量，并且不考虑变量之间的相互作用。但是该模型能够很好地用于判断哪些因素影响物种的分布。我们将在单元 5 中更详细地介绍地理模型和 Bioclim 模型。

统计回归模型同时需要物种分布有/无数据。正如我们在单元 3 中所学，分布无的数据可以是真实的分布无数据，也可以由构建的数据表示，我们称之为伪分布无数据。这些模型还使用环境数据，并且该算法使用所有数据来计算环境变量的系数，并且构建最能描述这些变量对物种分布影响的函数。统计回归模型可以处理连续和分类的预测因子，包括变量之间的交互作用。在单元 6 中，我们将讲解三种常用的物种分布模型的统计算法：广义线性模型，广义加性模型和多元自适应回归样条模型。

机器学习模型有很多不同的方法，所有的方法都使用环境数据。除了常用的 Maxent 方法使用分布有数据和背景数，大多数算法同时使用物种分布有和分布无数据。我们将在单元 5 中进一步介绍。各种机器学习模型都是基于决策树。在单元 7 中，我们将分类树的工作原理，了解更为复杂的基于分类树的模型：随机森林和增强回归树。我们将在本课程单元 7 中详细介绍另一种类型的机器学习模型：人工神经网络。

现在，主要的问题当然是如何选择物种分布模型的算法。这个问题没有简单的答案，因为这取决于很多不同的因素。虽然不能简单地推荐一种方法，我将简要概述各种模型的限制和假设，这可能会对你设计物种分布模型提供指导。

首先，你已有或想用的数据可能会限制你的某些选择。如果你没有任何环境数据，那么你只能使用地理模型，其结果仅仅指示一个物种的分布范围。如果你有环境条件数据，那么你可以设计一个真正的物种分布模型。下一步是查看你能得到的物种

数据。如果你只有分布有数据，你可以选择简单的框架模型，如 Bioclim。只有分布有数据的另一个可选方案，是 Maxent，这是一个分布有-背景模型，对比物种存在位点与所有可能位点的环境条件。另一个方案是具有真实的分布无数据或者伪分布无数据，选择同时考虑分布有/无数据的模型。可以是统计回归模型或机器学习模型。根据输入的数据，每个算法都有自己的假设和限制。如上所述，框架模型不能处理分类的预测变量和变量的交互作用。预测结果相对分布有/无数据模型和分布有-背景模型较差。与机器学习模型相比，统计模型往往对异常值和缺失值更为敏感。但机器学习模型可能过度拟合数据。我们将在单元 5 中看到，Maxent 有一个内置过程来避免过度拟合。但是，机器学习模型的优点在于它们能够处理海量数据。但是，如果你没有大量物种分布点，Maxent 或统计模型可能会更适用。请记住，这个指导并具有排他性，通常建议运行多个模型并比较其结果。

除了考虑输入数据对模型适用性的影响之外，模型的选择也取决于用户想要的结果。首先，不同模型结果的解释是不同的。Maxent 和 Bioclim 从环境角度进行建模，并预测物种分布的环境的适宜性。统计和机器学习模型则从物种角度出发，测试其在特定环境条件的地点的分布概率。另一个角度是用户的专业背景。虽然一些工具可能会使设计物种分布模型变得更加容易，但用户必须了解正在建模的内容。一些模型可能结果非常好，但理解和解释起来很复杂。另外，一些模型需要通过根据数据集，来设置选项为特定值。在这些情况下，运行具有默认配置选项的算法可能无法给出最佳结果。不是每个人都有时间或资源来学习新技术，因此你必须考虑你能做什么。我鼓励你探索物种分布模型，但需要记住，这是一个这是一个需要时间和投入才能充分理解的复杂问题。最后，还有一些实际的事情要记住，例如是否有可以免费获得的建模工具，你可能需要运行大数据和进行可视化输出，是否有使用相应的计算平台的权限。

除了所有这些选择之外，最后一件我想提到的事情是关于伪分布无数据。因为这个数据是由模型生成的，并不代表真实的观测结果。它可能会给模型引入某种错误。

因此，仔细考虑如何生成这些数据非常重要，具体包括两个方面：生成点的数量以及使用的方法。研究人员提供了通用指导原则，对统计模型而言，建议在研究区域随机生成 10,000 个伪分布位点，对机器学习模型来说，是在异于物种分布点环境条件的区域生成相同数量的伪分布无位点。再次提醒，建议你谨慎使用这一指南，并研究你感兴趣的算法的最新发展和建议。

没有一个算法能完美解决你所有的研究问题，你可能会失望但是所有这些选择都可以让你设计适合你的物种和研究区的物种分布模型。只是要确保你考虑了每个模型的标准和假设，并能解释你选择特定算法的理由。此外，在诸如 BCCVL 等工具中，你可以轻松地运行多种算法并比较其结果。因此，如果你不确定哪种算法最适合你的数据，可以选择多个算法。如果你使用多个模型，你可能得到的结果会稍微不同，在得出关于你感兴趣的物种分布的最终结果之前，考虑所有结果，总是很好。

我希望这个单元可以让你更好地了解设计物种分布模型时需要考虑的因素。在接下来的三个单元中，我们将介绍特定的算法细节。到时见！